说 明 书

基于大模型的问答方法、计算机系统、装置及存储介质

技术领域

本说明书一个或多个实施例涉及大模型领域，特别涉及一种基于大模型的问答方法、计算机系统、装置及存储介质。

背景技术

随着人工智能技术的不断进步，逐渐发展出大语言模型，为人们的工作和生活带来诸多便利。大语言模型能够理解和生成自然语言，完成翻译、问答、摘要生成、文本创作等多种任务。其中，基于大语言模型的问答具有较为广泛的应用，例如，用户可以对大语言模型进行提问，大语言模型可以根据用户的提问得到回答，并将回答输出给用户。目前需要一种针对指定知识文档基于大模型的问答方法。

发明内容

本说明书一个或多个实施例提供一种基于大模型的问答方法、计算机系统、装置及存储介质。

根据第一方面，提供一种基于大模型的问答方法，所述方法包括，基于预先根据指定知识文档构建的知识库，处理针对所述指定知识文档的提问信息，得到传输给目标大模型的提示信息，使所述目标大模型基于所述提示信息输出回答信息；获取基于所述回答信息确定的针对所述指定知识文档的更新数据；根据所述更新数据，对所述知识库进行更新。

根据第二方面，提供一种计算机系统，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如下步骤：基于预先根据指定知识文档构建的知识库，处理针对所述指定知识文档的提问信息，得到传输给目标大模型的提示信息，使所述目标大模型基于所述提示信息输出回答信息；获取基于所述回答信息确定的针对所述指定知识文档的更新数据；根据所述更新数据，对所述知识库进行更新。

根据第三方面，提供一种基于大模型的问答装置，装置方法包括，处理模块，用于基于预先根据指定知识文档构建的知识库，处理针对所述指定知识文档的提问信息，得到传输给目标大模型的提示信息，使所述目标大模型基于所述提示信息输出回答信息；获取模块，用

于获取基于所述回答信息确定的针对所述指定知识文档的更新数据；更新模块，用于根据所述更新数据，对所述知识库进行更新。

根据第四方面，提供一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述第一方面中任一项所述的方法。

本说明书的实施例提供的技术方案可以包括以下有益效果：

本说明书的实施例提供的基于大模型的问答方案，通过基于预先根据指定知识文档构建的知识库，处理针对指定知识文档的提问信息，得到传输给目标大模型的提示信息，使目标大模型基于提示信息输出回答信息，获取基于回答信息确定的针对指定知识文档的更新数据，根据更新数据，对知识库进行更新。从而能够在无需对大语言模型进行训练的前提下，保证了对知识库的更新优化，提高了问题回答的准确率。

应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本申请。

附图说明

为了更清楚地说明本说明书实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本说明书中记载的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图1是本说明书根据一示例性实施例示出的一种针对特定技术文档基于大模型进行提问的系统架构图；

图2是本说明书根据一示例性实施例示出的一种基于大模型的问答方法的流程图；

图3是本说明书根据一示例性实施例示出的一种基于大模型的问答装置框图；

图4是本说明书根据一示例性实施例示出的一种计算机系统框图。

具体实施方式

为了使本技术领域的人员更好地理解本说明书中的技术方案，下面将结合本说明书实施例中的附图，对本说明书实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本说明书一部分实施例，而不是全部的实施例。基于本说明书中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都应当属于本说明书保护的范围。

一方面，随着人工智能技术的不断进步，逐渐发展出大语言模型，为人们的工作和生活

带来诸多便利。大语言模型能够理解和生成自然语言，完成翻译、问答、摘要生成、文本创作等多种任务。其中，基于大语言模型的问答具有较为广泛的应用，例如，用户可以对大语言模型进行提问，大语言模型可以根据用户的提问得到回答，并将回答输出给用户。

另一方面，在一些场景下，例如，机构A开发出一套新的系统B，并配有相关的技术文档，系统B的用户需要通过查看相关的技术文档，了解系统B的操作使用规范以及技术介绍等信息。然而，由于技术文档可能信息量较大、内容较多，用户无法快速聚焦于技术文档中某个具体的问题。在相关技术中，通常由专人负责回答用户提出的关于技术文档中的问题，这就需要消耗人力的资源，并且，由人工查询并回答问题的效率一般也比较低，得到的回答准确率也无法得到保证。

因此，可以想到将大语言模型技术应用到该场景下，利用预先训练的大语言模型，帮助用户了解技术文档中的细节内容。例如，可以由用户向大语言模型提出针对技术文档的提问，由大语言模型对提问进行回答。然而，对于一些涉及领域较窄专业性较强的技术文档，通用大语言模型可能难以正确理解技术文档，如果需要利用大语言模型回答针对技术文档的提问，最好预先利用基于技术文档得到的语料对大语言模型进行训练。而训练大语言模型的各项成本都很高，也较为复杂。尤其是在技术文档需要经常更新的的情况下，利用大语言模型帮助用户了解技术文档的方式更是难以实施落地。

本说明书实施例提供了一种基于大模型的问答方案，通过基于预先根据指定知识文档（例如特定技术产品的相关技术文档等）构建的知识库，处理针对指定知识文档的提问信息，得到传输给目标大模型的提示信息，使目标大模型基于提示信息输出回答信息，获取基于回答信息确定的针对指定知识文档的更新数据，根据该更新数据，对知识库进行更新。从而能够在无需对大语言模型进行训练的前提下，保证了对知识库的更新优化，提高了问题回答的准确率。

图1是根据一示例性实施例示出的一种针对特定技术文档基于大模型进行提问的系统架构图。如图1所示，终端设备101为用户C持有的终端设备，用户C可以通过终端设备101登录技术文档D1对应的系统使用页面。技术文档D1对应的系统使用页面中嵌入有和辅助平台102连接的接口1，用户C可以通过接口1输入针对技术文档D1的提问信息，接口1可以将该提问信息传输给辅助平台102。

辅助平台102可以是针对特定的技术文档搭建的平台，辅助平台102中可以包括处理部分103和存储部分104。其中，存储部分104中至少存储有技术文档D1对应的数据，技术文档D1对应的数据可以包括文档内容信息、文档关键词信息以及专用术语及其解释等信息。处理部分103可以基于存储部分104中存储的技术文档D1对应的数据，对提问信息进行处理，得到

提示信息，并将提示信息传输给大模型105。

大模型105可以是市面上通用的LLM(Large Language Model，大语言模型)，大模型105可以根据辅助平台102传输的提示信息，得到回答信息，并将回答信息传输回辅助平台102。辅助平台102将回答信息返回给终端设备101，再由终端设备101将回答信息展示或播放给用户C。

另外，技术文档D1对应的系统使用页面中还嵌入有和运维平台106连接的接口2，用户C可以根据该回答信息，通过和运维平台106连接的接口2输入反馈信息，接口2将反馈信息传输给运维平台106。运维平台106可以将该反馈信息展示给运维人员，由运维人员根据该反馈信息确定存储部分104中需要更新的更新数据，并将更新数据输入至运维平台106。再由运维平台106将更新数据传输至辅助平台102，以对存储部分104存储的数据进行更新。

除此之外，在对辅助平台102中的处理部分103或存储部分104进行更新之前，或者在大模型105进行版本升级之前，可以由运维平台106提供预先构建的参考问答库。可以基于该参考问答库，检验对辅助平台102中的处理部分103或存储部分104进行更新的数据，或者新版本的大模型，是否会降低整个系统的性能。如果不降低整个系统的性能，则可以完成更新。

以下以一个具体应用实例，对通过上述系统进行提问的过程进行详细说明。具体来说，首先，用户C通过终端设备101登录技术文档D1对应的系统使用页面，并通过该系统使用页面中的接口1输入提问信息，该提问信息可以是一句通过自然语言表达的问题。例如，该提问信息可以是“在abc协议下如何传输信息”（abc为新产品中一种新的协议的专用名称）。可以通过接口1将该提问信息传输给辅助平台102。

接着，辅助平台102包括的处理部分103可以先通过语义识别的算法，对该提问信息进行语义识别，得到一组由语义关键词构成的词序列。并从辅助平台102包括的存储部分104中获取预先针对技术文档D1构建的术语列表，该术语列表中记录有技术文档D1所涉及的各个专用术语，以及各个专用术语对应的解释信息。处理部分103可以检测词序列中是否包括该术语列表中记录的专用术语，如果包括专用术语，则可以利用专用术语对应的解释信息对词序列进行更新操作。例如，处理部分103检测到词序列中包括该术语列表中记录的专用术语abc，并从该术语列表中获取专用术语abc对应的解释信息。接着，可以利用专用术语abc对应的解释信息，对词序列中的abc进行更新，例如，将专用术语abc对应的解释信息加入到词序列中的abc之后，得到新的词序列。

然后，处理部分103可以对词序列进行风险检测，识别词序列中是否包括敏感词。具体地，处理部分103可以先利用本地部署的风险检测模块对词序列进行风险检测，在检测通过的情况下，通过外部接口，利用国际通用的风险检测工具进行进一步的风险检测。

在成功通过两层风险检测的情况下，处理部分103可以从存储部分104中获取技术文档D1对应的文档内容信息以及文档关键词信息，并基于该文档关键词信息，从文档内容信息中查找与词序列中包括的语义关键词相关的部分文档内容。具体来说，文档关键词信息可以是基于预先从技术文档D1的文档内容中抓取的多个关键词得到的。处理部分103可以采用预设的相似度算法，从文档关键词信息中查找到与词序列中包括的语义关键词相关的文档关键词，并根据该相关的文档关键词，从文档内容信息中获取与提问信息相关的部分文档内容。可以根据该部分文档内容，生成提示信息。该提示信息中可以包括一些对回答的要求，上述查找到的部分文档内容以及经过标记处理的提问信息。例如，该提示信息中可以包括，abc协议的内容，在字符间添加了预设形式分隔符的提示信息“在-a-b-c-协-议-下-如-何-传-输-信-息”，以及需要大模型105在回答问题时遵守的规则。

处理部分103可以将生成的提示信息，传输给大模型105，由大模型105根据该提示信息进行分析，得到针对提问信息的回答信息，并将回答信息传输回处理部分103。处理部分103可以进一步对回答信息进行风险检测，例如可以通过本地部署的风险检测模块和/或外部通用的风险检测工具，对回答信息进行风险检测。如果风险检测通过，则处理部分103可以将该回答信息返回给终端设备101，终端设备101将该回答信息显示于技术文档D1对应的系统使用页面中以供用户C查看。

另外，用户C查看该回答信息之后，可以通过上述系统使用页面中的接口2，输入针对该回答信息的反馈信息。接口2可以将该反馈信息传输至运维平台106，其中，该反馈信息可以是用户C对该回答信息的意见等，诸如点赞或点踩信息，评价信息，或者修正意见信息等。例如，用户C觉得上述abc协议的定义内容不清楚，有歧义，可以将意见以反馈信息的形式输入接口2，并传输给运维平台106。运维平台106可以通过显示设备将该反馈信息展示给运维人员，运维人员可以参考该反馈信息，确定存储部分104中需要更新的更新数据，该更新数据可以是技术文档D1对应的文档内容等。例如，运维人员可以参考该反馈信息，重新编辑技术文档D1中对abc协议的定义的内容，作为对技术文档D1的更新数据。

在对辅助平台102进行更新（例如对处理部分103进行升级，或者对存储部分104中技术文档D1对应的文档内容进行更新）时，运维平台106可以进行更新测试。具体来说，运维平台106可以先获取预先构建的参考问答库，参考问答库中包括预先准备的多组问答信息，每组问答信息包括一个问题和一个针对该问题的答案。运维平台106首先可以从参考问答库中获取部分或全部问题，然后，将该部分或全部问题通过经过更新的辅助平台102，传输给大模型105。获取大模型105输出的各个问题的答案，作为测试答案，并从参考问答库中获取该部分或全部问题对应的答案，作为参考答案。可以将测试答案和参考答案进行比对，若比对

的结果指示测试答案和参考答案差异小于等于预设差异，说明更新测试通过，可以完成对辅助平台102的更新。若比对的结果指示测试答案和参考答案差异大于预设差异，说明更新测试未通过，可以将辅助平台102回滚到更新之前的状态。同时，可以重新设置对辅助平台102的更新数据，以重新对辅助平台102进行更新。

需要说明的是，在对大模型105进行版本升级时，运维平台106也可以进行上述更新测试。若更新测试通过，则可以完成对大模型105的版本升级，若更新测试未通过，则可以采用前一个版本的大模型。

下面将结合具体的实施例对本说明书提供的方案进行详细描述。

如图2所示，图2是根据一示例性实施例示出的一种基于大模型的问答方法的流程图，该方法可以应用于服务平台中。该服务平台可以实现为任何具有计算、处理能力的设备、平台、服务器或设备集群。该方法包括以下步骤：

在步骤201中，基于预先根据指定知识文档构建的知识库，处理针对指定知识文档的提问信息得到提示信息，并将提示信息传输给目标大模型。

在本实施例中，可以预先根据指定知识文档构建的知识库，其中，指定知识文档可以是针对特定产品或者专业领域的技术文档。例如，指定知识文档可以是开发的新设备或新系统的技术介绍文档，也可以是针对一个独立领域是的知识文档等，本实施例对指定知识文档的具体内容方面不限定。在获取到指定知识文档之后，可以从指定知识文档中抓取一些关键词，并基于这些关键词生成词向量。同时，还可以将指定知识文档切分成多个分片，每个分片中包括指定知识文档的部分文档内容，并且每个分片对应于一个或多个词向量。可以将上述词向量和上述多个分片存储为指定知识文档构建的知识库。

在本实施例中，服务平台可以根据上述预先构建的知识库，处理针对指定知识文档的提问信息，得到传输给目标大模型的提示信息。可选地，在处理提问信息之前，还可以对提问信息进行风险检测，在该提问信息中未包括敏感词的情况下，再执行处理提问信息的操作。例如，可以采用本地部署的风险检测对提问信息进行风险检测，也可以通过外部接口，利用外部提供的通用风险检测工具对提问信息进行风险检测等，本实施例对此方面不限定。在风险检测通过之后，可以执行处理提问信息的操作。

在处理提问信息的过程中，可以先通过预设的语义识别算法，对用户输入的提问信息进行语义识别，得到由语义关键词构成的目标词序列。该目标词序列例如可以是语义关键词按照其在提问信息中所对应的先后顺序，进行排列得到的序列。在获取到目标词序列之后，可以直接基于目标词序列生成提示信息，也可以先对目标词序列进行术语检测，根据检测结果对目标词序列进行处理之后，再生成提示信息。

具体地，可以通过如下方式对目标词序列进行术语检测：可以预先获取指定知识文档中的至少一个专用术语，生成术语列表，并在术语列表中记录每个专用术语在指定知识文档中所对应的解释信息，将术语列表与指定知识文档的知识库进行关联地存储。在进行术语检测时，首先，获取预先存储的术语列表，然后检测目标词序列中是否包括术语列表中的专用术语。如果目标词序列中不包括术语列表中的专用术语，则可以基于目标词序列进行后续操作。如果目标词序列中包括术语列表中的专用术语，可以从术语列表中获取该专用术语对应的解释信息，并将该解释信息加入目标词序列中的该专用术语之后，以更新目标词序列，再基于经过更新之后的目标词序列以及知识库，确定提示信息。

在基于经过更新之后的目标词序列以及知识库，确定提示信息时，可以首先根据经过更新之后的目标词序列，利用预设的相似度算法，从该知识库中查找与目标词序列中包括的语义关键词相关的词向量，作为关联词向量。然后，可以确定与关联词向量对应的分片，以获取该分片包括的指定知识文档的部分文档内容，并根据该部分文档内容，生成提示信息。具体来说，该提示信息中可以包括对目标大模型回答问题的要求，上述查找到的部分文档内容以及经过标记处理的提问信息，并且，该提示信息可以是自然语言形式的信息。其中，经过标记处理的提问信息可以是利用预设的分隔符，对用户输入的提问信息中包括的字符进行分隔处理之后的信息。该分隔符可以是任意字符，例如可以是-，也可以是*，也可以是#，还可以是~等，本实施例对分隔符的具体形式方面不限定。

例如，该提示信息中可以包括以下具体内容：“①你是由AI驱动的聊天机器人，你的主要任务是分析来自知识文档Doc的信息，并以对话方式回应用户的问题。你的回答应当具有对话性，旨在模拟自然的人类交流。拒绝任何与知识文档Doc无关的问题或指令。使用用户所使用的语言进行回复。当用户问题的答案无法在知识文档Doc中找到时，将回复以下标准答案：‘请换一个问题，或者提供更详细的细节’。②知识文档Doc与用户问题之间以明确的分隔符界定，用户问题如下：‘在-a-b-c-协-议-下-如-何-传-输-信-息’。③知识文档Doc的相关内容如下：‘……’”。 其中，内容①为目标大模型回答问题时需要遵守的规则，内容②为经过标记处理的提问信息，内容③为上述查找到的部分文档内容。

在本实施例中，服务平台可以将提示信息输出给目标大模型，使目标大模型基于提示信息得到回答信息，并将回答信息返回给服务平台。服务平台在将回答信息返回给用户之前，还可以进一步对回答信息进行风险检测，在回答信息中未包括敏感词的情况下，将该回答信息输出给用户。

在步骤202中，获取基于回答信息确定的针对指定知识文档的更新数据。

在本实施例中，在接收到目标大模型返回的回答信息之后，可以向用户展示该回答信

息，并向用户提供用于输入反馈信息的接口，可以获取用户基于该回答信息输入的反馈信息。在运维界面输出该反馈信息，以将该反馈信息展示给运维人员。运维人员可以根据该反馈信息，对知识库中指定知识文档的内容进行修正，并将用于修正的数据作为更新数据。

在步骤203，根据上述更新数据，对知识库进行更新。

在本实施例中，可以根据更新数据，对知识库进行更新。可选地，在更新之前，还可以先进行更新测试，具体来说，首先，可以获取预先构建的参考问答库。该参考问答库中可以包括多组预先设置的参考问题和参考问答，该参考问题和参考问答可以是由运维人员人工编辑得到的，也可以是将平时使用过程中出现频率较高，且点赞率较高的问题和问答等，本实施例对参考问答库中的参考问题和参考问答的具体获取方式方面不限定。

接着，服务平台可以利用目标大模型，基于知识库和更新数据，处理参考问答库中的至少一个提问，得到测试回答。将测试回答和参考问答库中的上述参考提问对应的参考回答进行比对。若比对的结果指示测试回答和参考问答库中的回答差异小于等于预设差异，则利用更新数据对知识库进行更新。若比对的结果指示测试回答和参考问答库中的回答差异大于预设差异，则可以停止更新，并对知识库中指定知识文档的内容进行进一步修正。

本说明书的上述实施例提供的基于大模型的问答方法，通过基于预先根据指定知识文档构建的知识库，处理针对指定知识文档的提问信息，得到传输给目标大模型的提示信息，使目标大模型基于提示信息输出回答信息，获取基于回答信息确定的针对指定知识文档的更新数据，根据更新数据，对知识库进行更新。从而能够在无需对大语言模型进行训练的前提下，保证了对知识库的更新优化，提高了问题回答的准确率。

应当注意，尽管在上述实施例中，以特定顺序描述了本说明书实施例的方法的操作，但是，这并非要求或者暗示必须按照该特定顺序来执行这些操作，或是必须执行全部所示的操作才能实现期望的结果。相反，流程图中描绘的步骤可以改变执行顺序。附加地或备选地，可以省略某些步骤，将多个步骤合并为一个步骤执行，和/或将一个步骤分解为多个步骤执行。

与前述基于大模型的问答方法实施例相对应，本说明书还提供了基于大模型的问答装置的实施例。

如图3所示，图3是本说明书根据一示例性实施例示出的一种基于大模型的问答装置框图，该装置可以包括：处理模块301，获取模块302和更新模块303。

其中，处理模块 301，用于基于预先根据指定知识文档构建的知识库，处理针对指定知识文档的提问信息，得到传输给目标大模型的提示信息，使目标大模型基于提示信息输出回答信息。

获取模块 302，用于获取基于回答信息确定的针对指定知识文档的更新数据。

更新模块303，用于根据更新数据，对知识库进行更新。

为了描述的方便，描述以上装置时以功能分为各种模块分别描述。当然，在实施本说明书时可以把各模块的功能在同一个或多个软件和/或硬件中实现。

The specification also provides a computer-readable non-transitory storage medium, wherein the storage medium stores a computer program, which, when executed by a processor, can be used to perform one or more steps of one or more methods described or illustrated herein or provides functionality described or illustrated herein. Herein, the computer-readable non-transitory storage medium or media may include one or more semiconductor-based or other integrated circuits (ICs) (such, as for example, field-programmable gate arrays (FPGAs) or application-specific ICs (ASICs)), hard disk drives (HDDs), hybrid hard drives (HHDs), optical discs, optical disc drives (ODDs), magneto-optical discs, magneto-optical drives, floppy diskettes, floppy disk drives (FDDs), magnetic tapes, solid-state drives (SSDs), RAM-drives, SECURE DIGITAL cards or drives, any other suitable computer-readable non-transitory storage media, or any suitable combination of two or more of these, where appropriate. A computer-readable non-transitory storage medium may be volatile, non-volatile, or a combination of volatile and non-volatile, where appropriate.

The specification also provides a computer system. FIG. 4 illustrates an example computer system 1000. In particular embodiments, one or more computer systems 1000 perform one or more steps of one or more methods described or illustrated herein. In particular embodiments, one or more computer systems 1000 provide functionality described or illustrated herein. In particular embodiments, software running on one or more computer systems 1000 performs one or more steps of one or more methods described or illustrated herein or provides functionality described or illustrated herein. Particular embodiments include one or more portions of one or more computer systems 1000. Herein, reference to a computer system may encompass a computing device, and vice versa, where appropriate. Moreover, reference to a computer system may encompass one or more computer systems, where appropriate.

This disclosure contemplates any suitable number of computer systems 1000. This disclosure contemplates computer system 1000 taking any suitable physical form. As example and not by way of limitation, computer system 1000 may be an embedded computer system, a system-on-chip (SOC), a single-board computer system (SBC) (such as, for example, a computer-on-module (COM) or system-on-module (SOM)), a desktop computer system, a laptop or notebook computer system, an interactive kiosk, a mainframe, a mesh of computer systems, a mobile telephone, a personal digital assistant (PDA), a server, a tablet computer system, or a combination of two or more of these. Where appropriate, computer system 1000 may include one or more computer

systems 1000; be unitary or distributed; span multiple locations; span multiple machines; span multiple data centers; or reside in a cloud, which may include one or more cloud components in one or more networks. Where appropriate, one or more computer systems 1000 may perform without substantial spatial or temporal limitation one or more steps of one or more methods described or illustrated herein. As an example and not by way of limitation, one or more computer systems 1000 may perform in real time or in batch mode one or more steps of one or more methods described or illustrated herein. One or more computer systems 1000 may perform at different times or at different locations one or more steps of one or more methods described or illustrated herein, where appropriate.

In particular embodiments, computer system 1000 includes a processor 1002, memory 1004, storage 1006, an input/output (I/O) interface 1008, a communication interface 1010, and a bus 1012. Although this disclosure describes and illustrates a particular computer system having a particular number of particular components in a particular arrangement, this disclosure contemplates any suitable computer system having any suitable number of any suitable components in any suitable arrangement.

In particular embodiments, processor 1002 includes hardware for executing instructions, such as those making up a computer program. As an example and not by way of limitation, to execute instructions, processor 1002 may retrieve (or fetch) the instructions from an internal register, an internal cache, memory 1004, or storage 1006; decode and execute them; and then write one or more results to an internal register, an internal cache, memory 1004, or storage 1006. In particular embodiments, processor 1002 may include one or more internal caches for data, instructions, or addresses. This disclosure contemplates processor 1002 including any suitable number of any suitable internal caches, where appropriate. As an example and not by way of limitation, processor 1002 may include one or more instruction caches, one or more data caches, and one or more translation lookaside buffers (TLB s). Instructions in the instruction caches may be copies of instructions in memory 1004 or storage 1006, and the instruction caches may speed up retrieval of those instructions by processor 1002. Data in the data caches may be copies of data of 80 in memory 1004 or storage 1006 for instructions executing at processor 1002 to operate on; the results of previous instructions executed at processor 1002 for access by subsequent instructions executing at processor 1002 or for writing to memory 1004 or storage 1006; or other suitable data. The data caches may speed up read or write operations by processor 1002. The TLBs may speed up virtual-address translation for processor 1002. In particular embodiments, processor 1002 may include one or more internal registers for data, instructions, or addresses. This disclosure contemplates processor 1002 including any suitable number of any suitable internal registers, where appropriate. Where appropriate, processor 1002 may include one or more arithmetic logic units (ALUs); be a

multi-core processor; or include one or more processors 1002. Although this disclosure describes and illustrates a particular processor, this disclosure contemplates any suitable processor.

In particular embodiments, memory 1004 includes main memory for storing instructions for processor 1002 to execute or data for processor 1002 to operate on. As an example and not by way of limitation, computer system 1000 may load instructions from storage 1006 or another source (such as, for example, another computer system 1000) to memory 1004. Processor 1002 may then load the instructions from memory 1004 to an internal register or internal cache. To execute the instructions, processor 1002 may retrieve the instructions from the internal register or internal cache and decode them. During or after execution of the instructions, processor 1002 may write one or more results (which may be intermediate or final results) to the internal register or internal cache. Processor 1002 may then write one or more of those results to memory 1004. In particular embodiments, processor 1002 executes only instructions in one or more internal registers or internal caches or in memory 1004 (as opposed to storage 1006 or elsewhere) and operates only on data in one or more internal registers or internal caches or in memory 1004 (as opposed to storage 1006 or elsewhere). One or more memory buses (which may each include an address bus and a data bus) may couple processor 1002 to memory 1004. Bus 1012 may include one or more memory buses, as described below. In particular embodiments, one or more memory management units (MMUs) reside between processor 1002 and memory 1004 and facilitate accesses to memory 1004 requested by processor 1002. In particular embodiments, memory 1004 includes random access memory (RAM). This RAM may be volatile memory, where appropriate. Where appropriate, this RAM may be dynamic RAM (DRAM) or static RAM (SRAM). Moreover, where appropriate, this RAM may be single-ported or multi-ported RAM. This disclosure contemplates any suitable RAM. Memory 1004 may include one or more memories 1004, where appropriate. Although this disclosure describes and illustrates particular memory, this disclosure contemplates any suitable memory.

In particular embodiments, storage 1006 includes mass storage for data or instructions. As an example and not by way of limitation, storage 1006 may include a hard disk drive (HDD), a floppy disk drive, flash memory, an optical disc, a magneto-optical disc, magnetic tape, or a Universal Serial Bus (USB) drive or a combination of two or more of these. Storage 1006 may include removable or non-removable (or fixed) media, where appropriate. Storage 1006 may be internal or external to computer system 1000, where appropriate. In particular embodiments, storage 1006 is non-volatile, solid-state memory. In particular embodiments, storage 1006 includes read-only memory (ROM). Where appropriate, this ROM may be mask-programmed ROM, programmable ROM (PROM), erasable PROM (EPROM), electrically erasable PROM (EEPROM), electrically alterable ROM (EAROM), or flash memory or a combination of two or more of these. This disclosure contemplates mass storage 1006 taking any suitable physical form. Storage 1006 may

include one or more storage control units facilitating communication between processor 1002 and storage 1006, where appropriate. Where appropriate, storage 1006 may include one or more storages 1006. Although this disclosure describes and illustrates particular storage, this disclosure contemplates any suitable storage.

In particular embodiments, I/O interface 1008 includes hardware, software, or both, providing one or more interfaces for communication between computer system 1000 and one or more I/O devices. Computer system 1000 may include one or more of these I/O devices, where appropriate. One or more of these I/O devices may enable communication between a person and computer system 1000. As an example and not by way of limitation, an I/O device may include a keyboard, keypad, microphone, monitor, mouse, printer, scanner, speaker, still camera, stylus, tablet, touch screen, trackball, video camera, another suitable I/O device or a combination of two or more of these. An I/O device may include one or more sensors. This disclosure contemplates any suitable I/O devices and any suitable I/O interfaces 1008 for them. Where appropriate, I/O interface 1008 may include one or more device or software drivers enabling processor 1002 to drive one or more of these I/O devices. I/O interface 1008 may include one or more I/O interfaces 1008, where appropriate. Although this disclosure describes and illustrates a particular I/O interface, this disclosure contemplates any suitable I/O interface.

In particular embodiments, communication interface 1010 includes hardware, software, or both providing one or more interfaces for communication (such as, for example, packet-based communication) between computer system 1000 and one or more other computer systems 1000 or one or more networks. As an example and not by way of limitation, communication interface 1010 may include a network interface controller (NIC) or network adapter for communicating with an Ethernet or other wire-based network or a wireless NIC (WNIC) or wireless adapter for communicating with a wireless network, such as a WI-FI network. This disclosure contemplates any suitable network and any suitable communication interface 1010 for it. As an example and not by way of limitation, computer system 1000 may communicate with an ad hoc network, a personal area network (PAN), a local area network (LAN), a wide area network (WAN), a metropolitan area network (MAN), or one or more portions of the Internet or a combination of two or more of these. One or more portions of one or more of these networks may be wired or wireless. As an example, computer system 1000 may communicate with a wireless PAN (WPAN) (such as, for example, a BLUETOOTH WPAN), a WI-FI network, a WI-MAX network, a cellular telephone network (such as, for example, a Global System for Mobile Communications (GSM) network), or other suitable wireless network or a combination of two or more of these. Computer system 1000 may include any suitable communication interface 1010 for any of these networks, where appropriate. Communication interface 1010 may include one or more communication interfaces 1010, where

appropriate. Although this disclosure describes and illustrates a particular communication interface, this disclosure contemplates any suitable communication interface.

In particular embodiments, bus 1012 includes hardware, software, or both coupling components of computer system 1000 to each other. As an example and not by way of limitation, bus 1012 may include an Accelerated Graphics Port (AGP) or other graphics bus, an Enhanced Industry Standard Architecture (EISA) bus, a front-side bus (FSB), a HYPERTRANSPORT (HT) interconnect, an Industry Standard Architecture (ISA) bus, an INFINIBAND interconnect, a low-pin-count (LPC) bus, a memory bus, a Micro Channel Architecture (MCA) bus, a Peripheral Component Interconnect (PCI) bus, a PCI-Express (PCIe) bus, a serial advanced technology attachment (SATA) bus, a Video Electronics Standards Association local (VLB) bus, or another suitable bus or a combination of two or more of these. Bus 1012 may include one or more buses 1012, where appropriate. Although this disclosure describes and illustrates a particular bus, this disclosure contemplates any suitable bus or interconnect.

还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

以上所述仅为本说明书的实施例而已，并不用于限制本说明书。对于本领域技术人员来说，本说明书可以有各种更改和变化。凡在本说明书的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。

权 利 要 求 书

1．一种基于大模型的问答方法，所述方法包括：

基于预先根据指定知识文档构建的知识库，处理针对所述指定知识文档的提问信息，得到传输给目标大模型的提示信息，使所述目标大模型基于所述提示信息输出回答信息；

获取基于所述回答信息确定的针对所述指定知识文档的更新数据；

根据所述更新数据，对所述知识库进行更新。

2．根据权利要求1所述的方法，其中，所述基于预先根据指定知识文档构建的知识库，处理针对所述指定知识文档的提问信息，得到传输给目标大模型的提示信息，包括：

对所述提问信息进行语义识别，得到由语义关键词构成的目标词序列；

对所述目标词序列进行术语检测；

若基于所述术语检测确定所述目标词序列中包括预设的专用术语，获取所述专用术语对应的解释信息，并将所述解释信息加入所述目标词序列中，以更新所述目标词序列；

基于经过所述更新之后的目标词序列以及所述知识库，确定所述提示信息。

3．根据权利要求2所述的方法，其中，所述对所述目标词序列进行术语检测，包括：

获取预先构建的术语列表；所述术语列表中包括至少一个专用术语以及每个专用术语对应的解释信息；

基于所述术语列表，对所述目标词序列进行术语检测；

其中，所述获取所述专用术语对应的解释信息包括，从所述术语列表中获取所述专用术语对应的解释信息。

4．根据权利要求3所述的方法，其中，所述基于经过所述更新之后的目标词序列以及所述知识库，确定所述提示信息，包括：

基于经过所述更新之后的目标词序列，从所述知识库中查找与所述目标词序列中包括的语义关键词相关的关联词向量；

基于所述关联词向量，确定所述指定知识文档中的部分文档内容；

根据所述部分文档内容，生成所述提示信息。

5．根据权利要求4所述的方法，其中，所述提示信息中包括通过预设方式对所述提问信息进行标记之后得到的信息。

6．根据权利要求1所述的方法，其中，在处理所述提问信息之前，所述方法还包括：

对所述提问信息进行风险检测；

在所述提问信息中未包括敏感词的情况下，执行处理所述提问信息的操作。

7．根据权利要求 1 所述的方法，其中，在所述目标大模型输出所述回答信息之后，所述方法还包括：

对所述回答信息进行风险检测；

在所述回答信息中未包括敏感词的情况下，将所述回答信息输出给用户。

8．根据权利要求 1 所述的方法，其中，所述获取基于所述回答信息确定的针对所述指定知识文档的更新数据，包括：

向用户展示所述回答信息，并获取用户基于所述回答信息输入的反馈信息；

在运维界面输出所述反馈信息，以获取针对所述指定知识文档的更新数据。

9．根据权利要求 1 所述的方法，所述根据所述更新数据，对所述知识库进行更新，包括：

获取预先构建的参考问答库；

利用所述目标大模型，基于所述知识库和所述更新数据，处理所述参考问答库中的提问，得到测试回答；

将所述测试回答和所述参考问答库中的回答进行比对；

若所述比对的结果指示所述测试回答和所述参考问答库中的回答差异小于等于预设差异，利用所述更新数据对所述知识库进行更新。

10、一种计算机系统，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如下步骤：

基于预先根据指定知识文档构建的知识库，处理针对所述指定知识文档的提问信息，得到传输给目标大模型的提示信息，使所述目标大模型基于所述提示信息输出回答信息；

获取基于所述回答信息确定的针对所述指定知识文档的更新数据；

根据所述更新数据，对所述知识库进行更新。

11．根据权利要求 10 所述的计算机系统，其中，所述基于预先根据指定知识文档构建的知识库，处理针对所述指定知识文档的提问信息，得到传输给目标大模型的提示信息，包括：

对所述提问信息进行语义识别，得到由语义关键词构成的目标词序列；

对所述目标词序列进行术语检测；

若基于所述术语检测确定所述目标词序列中包括预设的专用术语，获取所述专用术语对应的解释信息，并将所述解释信息加入所述目标词序列中，以更新所述目标词序列；

基于经过所述更新之后的目标词序列以及所述知识库，确定所述提示信息。

12．根据权利要求 11 所述的计算机系统，所述对所述目标词序列进行术语检测，包

括：

获取预先构建的术语列表；所述术语列表中包括至少一个专用术语以及每个专用术语对应的解释信息；

基于所述术语列表，对所述目标词序列进行术语检测；

其中，所述获取所述专用术语对应的解释信息包括，从所述术语列表中获取所述专用术语对应的解释信息。

13．根据权利要求 12 所述的计算机系统，其中，所述基于经过所述更新之后的目标词序列以及所述知识库，确定所述提示信息，包括：

基于经过所述更新之后的目标词序列，从所述知识库中查找与所述目标词序列中包括的语义关键词相关的关联词向量；

基于所述关联词向量，确定所述指定知识文档中的部分文档内容；

根据所述部分文档内容，生成所述提示信息。

14．根据权利要求 13 所述的计算机系统，其中，所述提示信息中包括通过预设方式对所述提问信息进行标记之后得到的信息。

15．根据权利要求 10 所述的计算机系统，所述处理器执行所述程序时还实现如下步骤：

在处理所述提问信息之前，对所述提问信息进行风险检测；

在所述提问信息中未包括敏感词的情况下，执行处理所述提问信息的操作。

16．根据权利要求 10 所述的计算机系统，所述处理器执行所述程序时还实现如下步骤：

在所述目标大模型输出所述回答信息之后，对所述回答信息进行风险检测；

在所述回答信息中未包括敏感词的情况下，将所述回答信息输出给用户。

17．根据权利要求 10 所述的计算机系统，其中，所述获取基于所述回答信息确定的针对所述指定知识文档的更新数据，包括：

向用户展示所述回答信息，并获取用户基于所述回答信息输入的反馈信息；

在运维界面输出所述反馈信息，以获取针对所述指定知识文档的更新数据。

18．根据权利要求 10 所述的计算机系统，其中，所述根据所述更新数据，对所述知识库进行更新，包括：

获取预先构建的参考问答库；

利用所述目标大模型，基于所述知识库和所述更新数据，处理所述参考问答库中的提问，得到测试回答；

将所述测试回答和所述参考问答库中的回答进行比对；

若所述比对的结果指示所述测试回答和所述参考问答库中的回答差异小于等于预设差异，利用所述更新数据对所述知识库进行更新。

19．一种基于大模型的问答装置，装置方法包括：

处理模块，用于基于预先根据指定知识文档构建的知识库，处理针对所述指定知识文档的提问信息，得到传输给目标大模型的提示信息，使所述目标大模型基于所述提示信息输出回答信息；

获取模块，用于获取基于所述回答信息确定的针对所述指定知识文档的更新数据；

更新模块，用于根据所述更新数据，对所述知识库进行更新。

20．一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令所述计算机执行权利要求 1-9 中任一项所述的方法。